

Summary

Summary

Methodological News features articles and developments relating to work done within the Methodology division. This issue contains four articles:

- Pufferfish differential privacy
- Improving prediction of dwelling occupancy for Census 2021
- Unsupervised machine learning for text analysis
- Data Innovation Hub

Pufferfish differential privacy

Pufferfish differential privacy

The ABS is conducting research to maximise data utility while providing state of the art privacy protection to data providers. Under the *Census and Statistics (Information Release and Access) Determination 2018*, the ABS provides passive confidentiality - which protects data providers who can demonstrate that the release of an ABS statistical output would be likely to allow their identification. These data providers are called passive claimants.

The current confidentiality method for protecting passive claimants is suppression; an aggregate statistic (e.g. a total) is not published if a passive claimant's value that contributes to the statistic is sensitive. One suppression leads to more suppressions to prevent the calculation of the original suppressed value based on related statistics. This limits the ABS's ability to meet an increasing user demand for more detailed business statistics. While suppression has performed well in the past, the ABS is continually striving to improve data confidentiality methods.

A privacy method that the ABS is currently investigating is **log-Laplace multiplicative perturbation**. This technique allows more detailed statistics to be safely published compared to suppression as it perturbs the passive claimant's value before it is used to produce aggregate statistics. Additional processes are not required to protect final statistical outputs. As a result, log-Laplace multiplicative perturbation is easier to implement than suppression, even when datasets and statistical outputs become complex. Furthermore, this approach protects against geospatial differencing risks where the passive claimant's value may be recovered from differencing aggregate outputs from overlapping regions.

Another advantage of the log-Laplace multiplicative perturbation is that it fits within a form of Pufferfish differential privacy (DP) framework. Our form of Pufferfish DP offers a privacy protection guarantee by connecting the $p\%$ rule with the Pufferfish DP framework. The $p\%$ rule is widely used at the ABS and other national statistical offices to determine if a passive claimant's value requires privacy protection. The $p\%$ rule is defined as follows; if a passive claimant's value can be estimated to within $p\%$ of its reported value, then it requires protection. In our form of Pufferfish DP, "secrets" are statements that take a form of "passive claimant A's reported value is within $p\%$ of the value x ". Log-Laplace multiplicative perturbation protects these "secrets" by ensuring users of our statistical outputs cannot confidently estimate a passive claimant's sensitive value to within $p\%$ of its reported value.

Preliminary results using Agriculture Statistics as a test case showed that log-Laplace multiplicative perturbation provides more data utility than suppression. Research is underway to explore its effectiveness with other datasets.

For more information, please contact Cedric Wong at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Improving prediction of dwelling occupancy for Census 2021

Improving prediction of dwelling occupancy for Census 2021

For some non-responding dwellings in the Census we do not have good information about whether they were occupied on Census night. For these dwellings a prediction about their occupancy status must be made. If they are deemed to be occupied then occupants will be imputed to them. In the 2016 Census these predictions were not based on a model and occupants were erroneously imputed to some non-responding dwellings. The Post Enumeration Survey adjusted for over- and under-count at broad levels. However, particularly in areas with large numbers of secure apartment buildings, these imputations contributed to inflated small area population counts.

For Census 2021 a prediction model has been developed to better classify non-responding dwellings to an occupancy status. Occupancy labels from the 2016 Census were used to train a model incorporating predictors from administrative data sources linked at the dwelling level. Key predictors were indicators of being on the Personal Income Tax, Medicare or Social Services datasets. Other useful predictors included area-level occupancy counts from the previous Census, dwelling type and geography. An extreme gradient boosting algorithm, implemented in R, was used and the models were trained, tuned and evaluated separately for each state.

We found these models were good for predicting long term unoccupied dwellings - holiday or rental homes being key examples. Prediction for dwellings that were short term unoccupied - i.e. where occupants were away just around the date of the Census - was less successful.

For an increasing number of regions Smartmeter electricity data is available and is a strong indicator of both long- and short-term dwelling occupancy. Due to privacy concerns this data was not used at the dwelling level. Instead, the Smartmeter data was used to derive a predicted count of occupied dwellings at an area level and these counts were used as an input to the models.

The predictions from the models take the form of probabilities of individual dwellings being occupied. In determining how best to use these probabilities two goals were considered: (i) classifying occupancy status for individual dwellings and (ii) obtaining accurate occupancy counts for small areas. The optimal approaches for these two goals were at odds. For (i) a probability cut-off value applied to each dwelling is optimal, but for (ii) the optimal approach would be to sum the predicted dwelling probabilities for the whole small area. A compromise using a stochastic process to generate an occupancy classification for each dwelling proportional to its predicted occupancy probability was chosen. This approach yielded expected values equivalent to the optimal approach for (ii) while being simple to implement within Census processing systems.

The prediction model and use of aggregate (area level) electricity data for determining dwelling occupancy status in the 2021 Census are still under investigation. However, the indications are that these models will provide a significant improvement to area level dwelling occupancy counts particularly for those areas with high numbers of secure apartment buildings. When combined with an improved method for imputation of occupants to non-responding dwellings, it is expected that these models will improve the quality of Census 2021 person counts.

For more information, please contact Sean Buttsworth at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Unsupervised machine learning for text analysis

Unsupervised machine learning for text analysis

In April 2020, the ABS trialled the application of unsupervised machine learning (ML) for the automatic analysis of free-text responses on two surveys relating to the impacts of COVID-19. The aim was to extract meaningful themes from these responses as quickly as possible.

Free-text is text or comments that users provide in response to open-ended questions. There is often no common structure or conceptual basis underlying the themes presented in free-text responses, and the themes cannot be anticipated prior to the survey. It is therefore difficult to automate the process of coding free-text responses, and manual coding is both resource-intensive and time consuming.

The first survey, Business Impacts of COVID-19, asked businesses if they had been adversely affected by COVID-19 (yes or no) and if so, how (free-text response). The second, the ABS Staff Wellbeing Survey, asked staff how they were currently feeling (respondents could select from range of very good to very bad); respondents were then provided an opportunity to explain why they were feeling this way, expressed in the form of a free-text response. In both surveys, analysis of the free-text responses required a quick turnaround (less than three weeks).

Text analysis involves three separate processes:

1. Preprocessing, where the text responses are cleaned, punctuation is removed, upper case letters

- converted to lower case, and words are reduced to their stems (e.g. 'working' is reduced to 'work')
2. Encoding (or embedding), in which the processed text is converted to a numerical form (such as a vector) that can be represented as a point in some space.
 3. Text is then grouped based on the distance between the points in the space. Depending on the encoding scheme used, similar text will be represented by points that are close together in the space. This step involves the use of Machine Learning (ML) to classify similar responses into groups or clusters.

Specifically, unsupervised ML can be used to group similar responses together. It is intended to work in situations where no pre-labelled data exists on which to train a model. Rather, the approach seeks to detect patterns/themes in the data without any pre-existing examples to learn from (in contrast with supervised approaches). Two unsupervised ML approaches were tested.

The first method used 'clustering': a method where similar text responses are grouped together into clusters. This method was tested on Business Impacts of COVID-19 Survey data and provided some insights into common effects of COVID-19 on businesses. Two predominant clusters were: 'staff working from home' and 'closure' of things.

The second method used probability distributions to identify major themes in text responses by identifying the most common words associated with a topic. This method was used to analyse the Staff Wellbeing Survey. Common themes identified when staff were feeling very good were: 'sun shining, good weather' and 'no commute time, sleep'. A common topic was identified for staff that were feeling very bad – this being 'personal pressure, children'.

Both methods were quick and processed hundreds of text responses in under ten minutes. While the methods were not perfect in grouping or clustering all text responses, they did work well to reduce the amount of manual analysis required. The clustering approach for Cycle 1 of the Business Impacts of COVID-19 Survey data successfully clustered about 30% of the free-text responses into coherent groups. The probability distribution method for the Staff Wellbeing Survey outperformed clustering analysis and successfully classified over 90% of free-text responses. Given further development, these methods show great potential for use in other ABS applications.

For more information, please contact Lisa-Maree Gulino at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Data Innovation Hub

Data Innovation Hub

In October 2020, the ABS established the Data Innovation Hub (DIH) initiative to help build enterprise capability in the innovative use of traditional and emerging data sources for official statistics. The DIH will bring together multidisciplinary teams of ABS methodologists, data scientists, technologists, and statistical subject matter experts to pilot new data solutions to improve the production of national statistics and provide new insights to government. Such solutions involve not only new data types and providers from across government, industry and academia, but also the methods and systems needed to systematically acquire, access, combine and analyse them.

The priorities, scope and objectives of innovation activity undertaken by the DIH will align with a set of Innovation Drivers that advance the strategic data capabilities of the ABS:

- Utilising non-traditional data sources – leveraging the information value of emerging 'big data' for statistical purposes, particularly in combination with existing administrative data sets.
- Delivering new information solutions – producing new analytical insights into current and emerging economic, social and environmental phenomena.
- Improving the statistical production process – automating or otherwise enhancing selection, coding, imputation, editing, and estimation processes.
- Combining heterogeneous multisource data – associating and resolving observations of statistical entities over multiple disparate data sets.
- Recognising patterns and correlations in data – identifying significant factors associated with a certain concept, and detecting outliers, anomalies, and spatial and temporal features in data.
- Protecting personal and sensitive data – applying privacy-preserving data transformations and models that minimise the loss of analytical utility.

The DIH will progress three significant projects in the first tranche of its work program to July 2021:

- Produce a monthly indicator of household final consumption expenditure (HFCE), using financial transactions data.
- Analyse changes in trade performance due to the disruption in global supply chains and the global

economic downturn caused by COVID-19, using merchandise trade records and business activity statements.

- Analyse the effect of reduced business activity due to COVID-19 on employment and jobs, using payroll records, business activity statements, social security payments and job advertisements.

A key objective of these projects is to quickly produce experimental outputs that demonstrate the value of new data sources and methods, and to set out a clear path to ongoing production. A second objective is to help build a technical environment that can support more effective and timely provision of information to policy makers.

For more information, please contact Ric Clarke at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

How to contact us and email subscriber list

To provide comments or feedback, or to be added to or removed from our electronic mailing list, please email methodology@abs.gov.au or contact:

Methodological News Editor
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
Belconnen ACT 2617

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

About this Release

Methodological News features articles and developments relating to work done within the Methodology division.

© Commonwealth of Australia

All data and other material produced by the Australian Bureau of Statistics (ABS) constitutes Commonwealth copyright administered by the ABS. The ABS reserves the right to set out the terms and conditions for the use of such material. Unless otherwise noted, all material on this website – except the ABS logo, the Commonwealth Coat of Arms, and any material protected by a trade mark – is licensed under a Creative Commons Attribution 2.5 Australia licence